SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

# Processing Big Data: Tools and Techniques Section 01

## CS 131

Spring 2024 3 Unit(s) 01/24/2024 to 05/13/2024 Modified 01/15/2024

## Contact Information

Ashish Khanchandani

Email: ashish.khanchandani@sjsu.edu

Office Hours: TuTh 11:00 AM to 12:00 PM Over Zoom

## Course Description and Requisites

In-depth study of essential tools and techniques for processing big data over the UNIX operating system and/or other operating systems. On UNIX, it includes using grep, sed, awk, join, and programming advanced shell scripts for manipulating big data.

Prerequisite(s): CS 46B or BIOL 123B with a grade of "C-" or better. Allowed Declared Majors: Computer Science BS, Data Science BS, MS Bioinformatics (MS BI).

Letter Graded

## Classroom Protocols

### Communication with the instructor

Students are requested to use the provided email to contact the instructor. The instructor does not write messages after normal business hours, on weekends or holidays. Reviewing code for the homework and technical trouble-shooting should be done during the office hours. Never send your entire code for an assignment to the instructor. The instructor will not fix all the bugs in your code.

Classroom Protocol

Course material developed by the instructor is the intellectual property of the instructor. Students can not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, hands-on exercises or homework solutions without instructor permission.

# Program Information

Diversity Statement - At SJSU, it is important to create a safe learning environment where we can explore, learn, and grow together. We strive to build a diverse, equitable, inclusive culture that values, encourages, and supports students from all backgrounds and experiences.

# Course Learning Outcomes (CLOs)

Upon successful completion of this course, students will be able to:

- Analyze, manipulate and process large-scale data with the UNIX/Linux command line and other operating systems.
- Develop shell scripts for use in data-intensive applications.
- Build data analysis pipelines, automate tasks, make analyses reproducible and shareable.
- Compare data analysis on the command line with use of graphical user interface and web-based tools.
- Solve big data challenges with the UNIX/Linux shell and command-line tools.
- Apply data science solutions to datasets from example domains, such as biology, business, finance.
- Perform big data analyses efficiently, document and reproduce analyses, use cloud computing for data- intensive problems.

# Course Materials

There are no required books for this class. All the necessary material will be available on the class Canvas web page.

The following resources are good for additional reference:

Textbooks:

- **Beginner**: UNIX Command Line: A Complete Introduction. William Shotts Jr
- **Moderate**: Linux Command Line and Shell Scripting Bible. Blum and Bresnahan
- **Advanced**: UNIX Power Tools. Jerry Peek, Tim O'Reilly, and Mike Loukides

Other good readings:

- Advanced Programming in the UNIX Environment. W. Richard Stevens, Stephen A. Rago. 3rd Edition, 2013, Addison-Wesley
- Introduction to UNIX and Linux. John Muster

- Data Science at the Command Line, 2nd Edition, Jeroen Janssens, Released August 2021, Publisher(s): O'Reilly Media, Inc. ISBN: 978149208791**Technology:**
- Practice of command-line operations will be done on IBM's computing cloud.
- Most assignments and worksheet tasks need to be submitted through Github. Details will be given in first assignment and worksheet instructions.

# Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on. The following information is subject to change with fair notice.

Reading assignments:

Readings may regularly be assigned for the next class. Students are expected to read the assigned materials.

Hands-On Worksheets:

We will have a number of hands-on worksheets. Please refer to Canvas for detailed instructions and deadlines. The worksheet submission page on Canvas closes after it is due. You need to submit the worksheets by their closing time on the due date. A worksheet will not be re-opened after its closing date. Late worksheets will not be accepted. As this is a fast-paced course, it is essential that you submit your worksheet homework in a timely fashion in order to keep up. 2 of 5 The purpose of the hands-on worksheets is to develop your understanding of the material and skills in using the command-line tools. The hands-on worksheets will involve learning how to use command line tools for analyzing and manipulating datasets from various domains, such as biology, business, finance. Students will use IBM's computing cloud and Amazon AWS for practice.

Homework assignments:

The assignments will be similar to worksheets. All assignments should be submitted on the corresponding assignment page in Canvas by 11:59 P.M. on the due date. Homework sent by email will not be graded, students need to upload them to Canvas. All homework solutions that students submit must be completely their own work. Solutions submitted on Canvas should reflect a student's own efforts. Do not write the code for anyone else. Never copy any code you find on another source, such as a website. Canvas automatically checks submissions for plagiarism from multiple online sources. Oral examination might be requested.

Examination Midterm exams:

There will be two midterm exams during the semester. Final exam: One final cumulative exam. The exams will contain multiple choice questions, true/false and short answer questions. No make-up exams except in case of verifiable emergency circumstances.

# ✓ Grading Information

## Criteria

Extra-credits assignments:

No extra-credit assignments are planned; However, the instructor may assign extra-credit assignments at his discretion with fair notice.

Late Submission (Applies only to assignments; not worksheets)

Late submissions within 24 hours will have 10% of the final grade deducted. Submissions over 24 hours late will have 20% grade of the grade deducted. Late submissions over 2 days will not be accepted unless prior consent has been granted by the instructor or in documented cases of emergency.

You can expect 3-5 Assignments including the mini project i.e your last assignment will be your mini project that you would present. 1 worksheet per week can be expected. The dates will be updated in the canvas.

Assignments : 30%

Worksheets : 20%

Presentation for the mini project : 10%

Midterms. : 20%

Finals : 20%

Note: This is subject to change but with fair notice

## Breakdown

| Grade | Range | Notes |
|---|---|---|
| A+ | 96% and above | |
| A | 92% to 95% | |

| Grade | Range | Notes |
|---|---|---|
| A- | 90% to 91% | |
| B+ | 87% to 89% | |
| B | 82% to 86% | |
| B- | 80% to 81% | |
| C+ | 77% to 79% | |
| C | 72% to 76% | |
| C- | 70% to 71% | |
| D+ | 67% to 69% | |
| D | 62% to 66% | |
| D- | 60% to 61% | |
| F | 59% and below | |

# University Policies

Per University Policy S16-9 (PDF) (http://www.sjsu.edu/senate/docs/S16-9.pdf), relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on the Syllabus Information (https://www.sjsu.edu/curriculum/courses/syllabus-info.php) web page. Make sure to visit this page to review and be aware of these university policies and resources.

# Course Schedule

Here's a breakdown of the course, lecture-by-lecture.

Note: This is a tentative schedule and is subject to change but with fair notice.

| When | Topic | Notes |
|---|---|---|
| 1/25 | Course Introduction | |
| 1/30 | Unix CLI Fundamentals | |
| 2/1 | Unix CLI Fundamentals Part 2 | |

| When | Topic | Notes |
|---|---|---|
| 2/6 | Files and Basic I/O with Unix | |
| 2/8 | Files and Basic I/O with Unix Part 2 | |
| 2/13 | File Permissions and Cron Jobs for Scheduling | |
| 2/15 | Directories, File structures and commands to work with Files | |
| 2/20 | Unix Processes Fundamentals and linking | |
| 2/22 | IPC (Inter Process Communication) : The why and how behind it. | |
| 2/27 | IPC (Inter Process Communication) Part 2 | |
| 2/29 | Tmux and Visualization | |
| 3/5 | Midterm Review and Discussion | |
| 3/7 | Midterm | |
| 3/12 | Text Parsing : sed, awk and regular expressions | |
| 3/14 | Text Parsing : sed, awk and regular expressions Part 2 | |
| 3/19 | Text Parsing : sed, awk and regular expressions Part 3 | |
| 3/21 | Introduction to Shell Scripting | |
| 3/26 | Shell Scripting Starter Kit | |
| 3/28 | Shell Scripting Constructs and Bringing it all together | |
| Spring break | | |
| 4/9 | Midterm 2 Review | |
| 4/11 | Midterm 2 | |
| 4/16 | Intro to Google Collab! | |
| 4/18 | Python Refresher | |
| 4/23 | Python Pandas Overview | |
| 4/25 | Plotting with Python | |
| 4/30 | Mini Project Presentations Session 1 | |

| When | Topic | Notes |
|---|---|---|
| 5/2 | Mini Project Presentations Session 2 | |
| 5/7 | Mini Project Presentations Session 3 | |
| 5/9 | Mini Project Presentations Session 4 and Final Review Session | |
| 5/17 | Finals | Friday, May 17 12:15-2:30 PM |